

Recovering 0 Kelvin Effective Hamiltonian Parameters from High-Temperature Disordered Phases

Elizabeth Decolvenaere and Michael J. Gordon

Department of Chemical Engineering, University of California Santa Barbara, Santa Barbara, California 93106, USA

Anton Van der Ven*

Materials Department, University of California Santa Barbara, Santa Barbara, California 93106, USA

(Dated: September 26, 2016)

Effective Hamiltonians, when used in tandem with statistical mechanics techniques, offer a rigorous connection between 0 Kelvin *ab-initio* predictions and finite temperature experimental observations. For alloys, cluster expansion Hamiltonians can coarse-grain out the complex, many-body electron problem of density functional theory, yielding a series of simple site-wise basis functions (e.g., products of site occupancy variables) on an atomic scale. The resulting energy polynomial is computationally inexpensive, and hence suitable for the (tens of) thousands of calculations of large systems required by stochastic methods. We present a new method to run the statical mechanics problem “in reverse”, using high-temperature observations and thermodynamic connections to construct an effective Hamiltonian and thereby predict the 0 Kelvin energy spectrum and associated ground states. By re-examining the cluster expansion formalism through the lens of entropy-maximization approaches, we develop an algorithm to select clusters and determine cluster interactions using only a few, high-temperature experiments on disordered phases. We demonstrate that our approach can recover not only the stable ground states at 0 Kelvin, but also the full phase behavior for three realistic two-dimensional and three-dimensional alloy test-cases.

I. INTRODUCTION

First-principles electronic structure methods, such as density functional theory (DFT), can provide a unique view into atomic-scale properties that are otherwise *inaccessible* via experiment. Statistical mechanics, or other scale bridging techniques, can then connect the quantum mechanical energy spectrum to the realm of experimentally observable, and industrially-relevant, temperatures and length scales. Directly utilizing first-principles electronic structure methods in statistical mechanics schemes (e.g., to calculate the energy of every microstate), though, is in general computationally intractable. While *ab-initio* molecular dynamics¹ is increasingly being used to probe high temperature behavior^{2,3}, it remains restricted to artificially small periodic unit cells and short simulation times^{4,5}. Instead, atomistic models^{5–21} are more often used to represent a first-principles landscape as a function of relevant degrees of freedom. The path from electronic structure to the laboratory, however, is almost entirely one-way: should an *ab-initio* method prove unreliable when compared to experiment, the experiments cannot be meaningfully used to inform and improve the electronic structure model with the same detail and precision as a direct, first-principles method.

This situation motivates the development of a technique that goes “in reverse”, whereby measurements of an easily accessible, high-temperature, and disordered phase are used to develop an atomistic model that is accurate at zero Kelvin. The advantages of such an approach are many. An accurate atomistic model parameterized with high temperature data can be used to predict the energy spectrum over the microstates of a solid,

as well as to reveal thermodynamic ground states that are otherwise difficult to determine experimentally (e.g., due to sluggish kinetics at low temperatures)²². Furthermore, the model can be applied in conventional Monte Carlo simulations to predict the full phase diagram^{23–26}, or with variance constrained Monte Carlo to predict free energies inaccessible to experiment (e.g., inside the spinodal of a miscibility gap)²⁷. Even kinetic properties, such as diffusion^{28–30} and precipitate nucleation and growth³¹, can be elucidated with such a model.

Effective Hamiltonians^{7,9,14,16,32–34}, which have seen extensive use in the literature^{19,23–26,28,30,35–45}, provide a framework well-suited to developing such a model. They have proven to be powerful tools to extrapolate first-principles energy landscapes and come in many forms. A harmonic Hamiltonian expressed in terms of inter-atomic force-constants, for example, extrapolates first-principles force-displacement relations to predict phonon properties and vibrational free energies. Cluster expansions^{7,9,14} and anharmonic lattice dynamics Hamiltonians^{10,11,18–20} have enabled the first-principles study of alloy phase diagrams and structural phase transitions with Monte Carlo^{10,12,18,26,29,45,46}. In their most rigorous form, an effective Hamiltonian can be formulated as a linear expansion in a set of basis functions, expressed in terms of variables that describe particular atomic degrees of freedom. Alloy Hamiltonians, for example, commonly referred to as “cluster expansions”, are expressed in terms of polynomials of occupation variables associated with clusters of sites (e.g., pairs, triplets etc.) in the crystal^{7,9,14}. The resulting polynomial is computationally inexpensive, and thus well-suited for stochastic methods such as Monte Carlo, which require tens of thousands of energy evaluations to calculate accurate thermodynamic properties.

Here, we explore the possibility of developing an “experiments-first” effective Hamiltonian, using high temperature experiments to predict zero Kelvin behavior. We present a new method of parametrizing the Hamiltonian using experimental data of the disordered state instead of zero Kelvin quantum mechanical predictions. The approach not only yields a parameterization of the expansion coefficients, but also suggests the most probable truncation of the Hamiltonian. Overall, the method enables the construction of an accurate atomistic model of crystalline materials suitable for a wide variety of stochastic simulation techniques. Our approach provides a new tool to develop full phase diagrams and probe otherwise difficult-to-measure thermodynamic properties, using only a small number of high-temperature observations of a disordered phase.

II. A THERMODYNAMIC APPROACH TO CLUSTER EXPANSION PARAMETERS

We illustrate our approach of parameterizing an effective Hamiltonian with high temperature experimental data in the context of a binary A-B crystalline alloy. The approach is, nevertheless, general, and can be applied to any effective Hamiltonian constructed as a linear expansion of basis functions that depend on one or more atomic degrees of freedom (e.g., local magnetic moments, atomic displacements etc.).

A. Overview of the Cluster Expansion Formalism

Each crystal site, i , of a binary solid is occupied by one of two components: A or B. We can assign an occupation variable to each site, $\sigma_i = \pm 1$ (e.g., A = +1 and B = -1), such that the arrangement of A and B atoms in a crystal of M sites is completely specified by $\bar{\sigma} = \{\sigma_1, \sigma_2, \dots, \sigma_M\}$. Each configuration has an occupation-dependent energy $E = E(\bar{\sigma})$ that can, in principle, be calculated with a first-principles electronic structure method. Sanchez *et al*^{7,9} showed that the configuration dependence of the energy of a crystal can be expanded in terms of an orthogonal basis of cluster functions $\phi_\delta(\bar{\sigma})$ defined as a product of occupation variables belonging to a cluster of sites in the crystal:

$$\phi_\delta(\bar{\sigma}) = \prod_{i \in \delta} \sigma_i, \quad (1)$$

where δ is a geometric *cluster* of sites. The energy $E(\bar{\sigma})$ can then be written as:

$$E(\bar{\sigma}) = \sum_{\delta \in L} \phi_\delta(\bar{\sigma}) V_\delta, \quad (2)$$

with the constant expansion coefficients, V_δ , capturing the many-body physics of the interactions among the

atoms (or molecules, or vacancies) of the crystal. The expansion coefficients are referred to as “effective cluster interactions” (ECIs).

Clusters (of sites) related under space group symmetry operations of the crystal (including translation) will have the same V_δ . Equation (2) can be simplified by collecting clusters with identical V_δ into groups Ω_α , where α represents a prototype of a particular orbit of symmetrically equivalent clusters (e.g., all nearest-neighbor pairs). Equation (2) then becomes:

$$E(\bar{\sigma}) = \sum_{\alpha} \Phi_{\alpha}(\bar{\sigma}) V_{\alpha}, \quad (3)$$

with $\Phi_{\alpha}(\bar{\sigma}) = \sum_{\delta \in \Omega_{\alpha}} \phi_{\delta}(\bar{\sigma})$. The sum in Equation (3) is restricted to symmetrically distinct clusters no larger than the volume of the crystal, with no more than M members. The Φ_{α} , which we will call *extensive cluster functions*, differ from the correlations $\varphi_{\alpha}(\bar{\sigma})$ conventionally defined in the literature⁷:

$$\varphi_{\alpha}(\bar{\sigma}) = \frac{\sum_{\delta \in \Omega_{\alpha}} \phi_{\delta}(\bar{\sigma})}{m_{\alpha} N_P}, \quad (4)$$

where N_P is the number of primitive unit cells in the crystal and m_{α} is the multiplicity of cluster α per primitive unit cell. Φ_{α} and φ_{α} are then related by a factor of $m_{\alpha} N_P$, such that Φ_{α} scales with the size of the crystal. This extensive property will prove useful in applying Legendre transforms to develop thermodynamic potentials for ensembles of fixed extensive cluster functions.

As a last step, it is convenient to express the cluster expansion of the configurational energy as a scalar product between two vectors: one vector being the collection of extensive cluster functions, $\bar{\Phi}(\bar{\sigma})$, and the other being the corresponding ECIs, \bar{V} . Equation (3) can then be expressed as a dot product:

$$E(\bar{\sigma}) = \bar{\Phi}(\bar{\sigma}) \cdot \bar{V} \quad (5)$$

Although formally rigorous, a challenge in making the cluster expansion practical is the determination of numerical values for the expansion coefficients \bar{V} . The traditional approach is to use DFT or one of its extensions to calculate the energy of a number of configurations, $\bar{E}(\bar{\sigma})_{\text{DFT}}$, and then inverting Equation (3) to determine the ECI using one of many schemes^{22,47–55}. The cluster expansion, however, extends over all cluster basis functions, which for a binary alloy having a crystal of M sites is equal to 2^M . The expansion must, therefore, be truncated. The choice of clusters to remain in the expansion is an often-studied problem with no simple solution⁵⁶. Previous work, though, suggests the set of clusters to be sparse^{23–26,28,30,35–39,41,42,44}, and many techniques have been developed to choose a few basis functions from a large pool of candidates^{50,53,57}.

B. Thermodynamic Relationships

Experiments are unable to provide direct access to the energies of individual microstates $\bar{\sigma}$. Methods for measuring internal energies or enthalpies return only the average over many microstates. Hence, a method based on an inversion of Equation (3), relying on experimental measurements, is unlikely to be found. However, by assigning a thermodynamic interpretation to the expansion coefficients, \bar{V} , of a cluster expansion, other expressions can be derived that relate averages of spatial correlations over clusters of sites, which can be measured with a variety of local or reciprocal probes, to the expansion coefficients \bar{V} .

It is convenient to work in the canonical ensemble (constant temperature T , number of sites M , and alloy concentration N_A), which has as partition function Z and free energy A :

$$Z(T, M, N_A) = \sum_{\bar{\sigma}} e^{-\frac{\Phi(\bar{\sigma}) \cdot \bar{V}}{k_b T}} \quad (6)$$

$$A(T, M, N_A) = -k_b T \ln [Z] \quad (7)$$

The sum is restricted to configurations $\bar{\sigma}$ having fixed composition, and k_b is the Boltzmann constant. Starting with the canonical free energy, we can produce a number of derivatives, some of which have been discussed in previous work⁵⁸. We highlight a few that are of practical importance here:

$$\frac{\partial A}{\partial V_\alpha} = \langle \Phi_\alpha \rangle \quad (8)$$

$$\frac{\partial^2 A}{\partial V_\alpha \partial V_\beta} = \frac{\partial \langle \Phi_\alpha \rangle}{\partial V_\beta} = \frac{\partial \langle \Phi_\beta \rangle}{\partial V_\alpha} = -\frac{\text{cov}[\Phi_\alpha, \Phi_\beta]}{k_b T} \quad (9)$$

$$\frac{\partial^2 A}{\partial V_\alpha \partial T} = \frac{\partial \langle \Phi_\alpha \rangle}{\partial T} = -\frac{\partial S}{\partial V_\alpha} = \frac{\text{cov}[\Phi_\alpha, (\bar{V} \cdot \Phi)]}{k_b T^2} \quad (10)$$

where $\langle y \rangle = \sum_{\bar{\sigma}} y \frac{\exp[-\frac{\bar{V} \cdot \Phi(\bar{\sigma})}{k_b T}]}{Z}$ denotes the ensemble average of y and $\text{cov}[y, z] = \langle yz \rangle - \langle y \rangle \langle z \rangle$ denotes the ensemble covariance of y and z . S in Equation (10) refers to the entropy.

Equations (9) and (10) are response functions, measuring how the ensemble average of an extensive cluster function, $\langle \Phi_\alpha \rangle$, responds to a change in either an ECI, V_β , or the temperature. Equation (10) is especially useful after expanding the covariance of the products and rearranging slightly:

$$k_b T^2 \frac{\partial \langle \Phi \rangle}{\partial T} = \mathbf{cov}[\bar{\Phi}, \bar{\Phi}] \cdot \bar{V} \quad (11)$$

with $\mathbf{cov}[\bar{\Phi}, \bar{\Phi}]$ denoting a matrix, with each element of this matrix, $(\mathbf{cov}[\bar{\Phi}, \bar{\Phi}])_{\alpha, \beta}$, corresponding to an ensemble averaged covariance between a pair of extensive clus-

ter functions, Φ_α and Φ_β . The left hand side of Equation (11) is a column vector of the temperature derivatives of the ensemble averages of the extensive cluster functions Φ_α , multiplied by $k_b T^2$.

Equation (11) is a crucial component of the approach as it provides a connection between a *measurable* set of variables, $\mathbf{cov}[\bar{\Phi}, \bar{\Phi}]$ and $k_b T^2 \frac{\partial \langle \Phi \rangle}{\partial T}$, and a desirable (but immeasurable) set of coefficients, \bar{V} . Once values have been measured for the temperature dependence of the extensive cluster functions, and for covariances between pairs of extensive cluster functions, it should in principle be possible to invert Equation (11) to recover the expansion coefficients \bar{V} . These expansion coefficients can then be used in standard statistical mechanics approaches to determine ground states and to calculate the *full* phase diagram. Hence, with only a few measurements, information about the entire phase space can be generated.

C. Entropy-Maximizing Basis Function Selection

While Equation (11) offers the potential to extract the ECI of a cluster expansion from experimental measurements of extensive cluster functions, Φ_α , and their covariances, a direct inversion is, in general, infeasible. Experience with first-principles parameterized cluster expansions shows that these Hamiltonians are typically sparse, converging rapidly as the cluster size of a basis function increases, both in spatial extent and number of sites. Even when the clusters are small, their corresponding ECI may be close to zero. Before Equation (11) can be inverted, it is therefore necessary to devise a method to determine the “correct” sparse set of clusters (i.e., non zero elements in \bar{V}). Biased regression schemes (such as l^1 -norm penalization⁵⁹), while attractive for DFT-based cluster expansions^{51,54,55}, perform poorly when elements in the design matrix (i.e., $\mathbf{cov}[\bar{\Phi}, \bar{\Phi}]$) are correlated⁶⁰. As the columns in our covariance matrix are themselves correlated, we require an external cluster-selection step that is robust to this feature.

To this end, we again rely on a thermodynamic interpretation of the expansion coefficients \bar{V} . The entropy-maximization approach of Jaynes⁶¹ (MAXENT) can be employed to develop a simple metric to judge whether a given cluster should be included or excluded in a final regression scheme to extract the non-zero \bar{V} from Equation (11). Treating the ECI as thermodynamic variables, we can re-cast the problem in the form of finding a set of parameters, \bar{V} , which satisfy:

$$\frac{\partial A}{\partial \bar{V}} = \langle \bar{\Phi} \rangle_{\text{obs}}, \quad (12)$$

where $\langle \bar{\Phi} \rangle_{\text{obs}}$ is an *observed* value of the extensive cluster functions. This relation yields a microstate distribution that maximizes the “information entropy” of the system, given the constraint that $\langle \bar{\Phi} \rangle$, the ensemble average, is equal to $\langle \bar{\Phi} \rangle_{\text{obs}}$. The Lagrange multipliers in this

constrained maximization problem are, conveniently, the ECIs. Our entropy-maximizing solution is then given by the *stationary points* with respect to \bar{V} of the free energy Υ :

$$\Upsilon(T, M, N_A, \bar{V}, \langle \bar{\Phi} \rangle_{\text{obs}}) = A(T, M, N_A) - \bar{V} \cdot \langle \bar{\Phi} \rangle_{\text{obs}}. \quad (13)$$

Finding the stationary points is as simple as solving $\frac{\partial \Upsilon}{\partial \bar{V}} = 0$ (which returns Equation (12)). When the $\langle \bar{\Phi} \rangle_{\text{obs}}$ are measured in a thermodynamically stable phase, these stationary points are *maxima*, as proven by the sign of the Hessian of A (and, therefore, of Υ) in \bar{V} :

$$\frac{\partial^2 \Upsilon}{\partial \bar{V}^2} = -\frac{\text{cov}[\bar{\Phi}, \bar{\Phi}]}{k_b T} \leq 0. \quad (14)$$

The strict seminegative-definite nature of the Hessian of Υ for thermodynamically stable phases guarantees a single maximum only. This means that any changes in \bar{V} that increase Υ are moving us towards that global maximum — there are no local maxima upon which to become trapped. Therefore, if we can evaluate how Υ changes when a cluster is included or excluded, we can use the sign of $\Delta \Upsilon$ to determine if that cluster is moving us towards or away from the MAXENT solution.

A difficulty with Equation (13) is that we do not know the free energy A of the phase in which the $\langle \bar{\Phi} \rangle_{\text{obs}}$ were measured. However, for a disordered solid solution, we can approximate it by performing a Taylor expansion of $A(\bar{V})$ around the non-interacting crystal ($\bar{V} = 0$) corresponding to an ideal solution. To first order:

$$\Upsilon(T, M, N_A, \bar{V}, \langle \bar{\Phi} \rangle_{\text{obs}}) \approx A_0 + \bar{V} \cdot (\langle \bar{\Phi} \rangle_0 - \langle \bar{\Phi} \rangle_{\text{obs}}) \quad (15)$$

where A_0 is the ideal solution free energy, and $\frac{\partial A}{\partial \bar{V}}|_{\bar{V}=0} = \langle \bar{\Phi} \rangle_0$ is the vector of ideal-solution extensive cluster functions, which can easily be evaluated, as the sites of any cluster in an ideal solution are uncorrelated by definition. With the Taylor expansion approximation to Υ , the criterion $\Delta_\alpha \Upsilon$ as to whether or not a cluster α should be included is then:

$$\Delta_\alpha \Upsilon(T, M, N_A, \langle \bar{\Phi} \rangle_{\text{obs}}) \approx (\langle \bar{\Phi} \rangle_0 - \langle \bar{\Phi} \rangle_{\text{obs}}) \cdot [\bar{V}_{\text{new}}(\langle \bar{\Phi} \rangle_{\text{obs}}) - \bar{V}_{\text{old}}(\langle \bar{\Phi} \rangle_{\text{obs}})] \quad (16)$$

where $\bar{V}_{\text{new}}(\langle \bar{\Phi} \rangle_{\text{obs}})$ and $\bar{V}_{\text{old}}(\langle \bar{\Phi} \rangle_{\text{obs}})$ refer to the values of the ECIs calculated using Equation (11) with cluster α included and excluded, respectively. By testing each candidate cluster α for $\Delta_\alpha \Upsilon > 0$, we can differentiate between relevant clusters with small ECIs, and clusters with 0 ECIs that recover nonzero values due to regression

error. This algorithm is described in Appendix A, and requires only a single pass through the set of all clusters. For reasons of numerical stability, only cluster observations on the same length-scale of any “selected” clusters are used for subsequent evaluations of Equation (11).

The uniqueness of the maximum of Υ is only guaranteed where the free energy varies smoothly, i.e., far from a phase boundary. Additionally, as our Taylor expansion is based around the ideal solution, observations should only be drawn from the disordered phase. This is an easy region to access experimentally, and agrees well with the goals outlined at the beginning of this section. Using Equation (16), we can determine the ideal set of clusters to include, and with Equation (11), we can solve for their ECIs. These clusters and ECIs are sparse, thermodynamically-consistent, share a one-to-one mapping with the observed extensive cluster functions, and can be found using only a few observations of the high-temperature, disordered phase.

III. TESTING THE HAMILTONIAN INVERSION APPROACH ON SIMULATED DATA

We used simulated data sets to test the viability of the methodology developed in Section II to parameterize an effective Hamiltonian to high temperature measurements. Benchmarking of the approach was performed on three binary systems (A-B alloys) with their configurational energy described by cluster expansion Hamiltonians. This included two systems on a 2D triangular lattice using: (I) only nearest and next-nearest neighbor (NN and NNN) interactions, and (II) six pseudo-random interactions, including three and four-body clusters. System I has been characterized in-depth by Glosli and Plischke⁶². We also studied a 3D FCC lattice (III) using clusters and ECIs generated from first-principles to model the Au-Cu system by Z. Lu, *et al*⁶³. For all systems, we report our results using the following dimensionless, reduced units:

$$\tau = \frac{k_b T}{V_{NN}} \quad m = \frac{\mu_A - \mu_B}{V_{NN}} \quad x_A = \frac{N_A}{M} \\ v_i = \frac{V_i}{V_{NN}} \quad e = \frac{E}{V_{NN}}$$

where V_{NN} is the nearest-neighbor-pair ECI from the *original* cluster expansion, and E and e refer to *any* type of energy, in the absolute and dimensionless units, respectively. An ECI that is strictly zero, i.e., $v_\alpha = 0$, is equivalent to cluster function ϕ_α being excluded from the cluster expansion.

The normalized chemical potential difference m is related to the slope of the alloy free energy as a function of alloy composition x_A . The reference states for the model cluster expansions were defined such that the energies for pure A and pure B are both equal to zero. With these reference states, very negative values of m correspond to B-rich alloys while very positive values of m correspond

to A rich alloys. Equi-composition alloys have intermediate values of m that are centered around zero.

The simulated data was generated with semi-grand canonical Monte Carlo simulations performed using the three model cluster expansions, using the CASM code^{19,24,29,64}. While the methodology developed in Section II relies upon derivatives taken at constant composition, x_A , rather than at constant chemical potential, m , switching from the canonical to the semi-grand canonical ensemble requires only minor modifications to the equations and changes none of the analysis⁶⁵. The Monte Carlo simulations were used to calculate ensemble averages of the extensive cluster functions, $\langle\Phi_\alpha\rangle$, and their covariances, $\text{cov}[\overline{\Phi}, \overline{\Phi}]$, in the disordered solid solution of the model alloys at high temperature. These two quantities represent the “experimental data” needed to invert Equation (11) to determine the ECIs. A 30×30 periodic supercell of the 2D triangular lattice was used to simulate data for systems I and II, while a $14 \times 14 \times 14$ periodic supercell of the FCC primitive cell was used to generate data for system (III). All measurements were taken from cooling runs at constant dimensionless chemical potential m . The number of passes (N_{pass}), starting dimensionless temperature (τ_0), incremental dimensionless temperature ($\Delta\tau$), and incremental dimensionless chemical potential (Δm) are given in Table (I):

For all three model alloys, we found a strong dependence of the recovered ECI on the value of m used to generate the simulated experimental data sets. Data sets collected at chemical potentials that stabilize B-rich alloys or A-rich alloys (i.e., very negative or very positive values of m) were less robust, as changes in $\langle\overline{\Phi}\rangle$ became small at near-pure compositions. However, in the chemical potential range that stabilizes a more equi-compositional alloy, a more consistent and reliable set of ECIs could be recovered (provided the values of m and τ were not too close to a phase transition). To compare the robustness of simulations performed at different values of m , we employed the following “consistency score” figure-of-merit:

$$S_m = \frac{2}{\|\overline{v}_m - \overline{v}_{m+\Delta m}\| + \|\overline{v}_m - \overline{v}_{m-\Delta m}\|}, \quad (17)$$

where \overline{v}_m is the vector of (reduced) ECIs evaluated for a simulation performed at chemical potential m , and Δm is the chemical potential step size used when performing multiple simulations. S_m corresponds to the (reciprocal of the) average Euclidean norm of ECIs evaluated at three chemical potentials m and $m \pm \Delta m$. This consistency score is used to evaluate when the inversion algorithm ceases to provide reliable results, due to divergences or zeros in $\text{cov}[\overline{\Phi}, \overline{\Phi}]$ at phase boundaries or compositional extremes. The reciprocal form provides easier interpretation of the results, and maps the “steadiest” solutions into the largest scores.

In the following three sections, we summarize the thermodynamic phase behavior of each model system and describe how the inverted cluster expansions compare to

the original cluster expansions used to generate the high temperature data sets. For each alloy system, we determined the final \overline{v} with the following process. High temperature measurements (averages of the extensive cluster functions, $\langle\Phi_\alpha\rangle$, and their covariances, $\text{cov}[\overline{\Phi}, \overline{\Phi}]$) were calculated for a range of m values. The temperature range was chosen to be both narrow and near (but not at) the highest-temperature phase transition. For each value of m , a sparse vector, \overline{v}_m was determined using the algorithm of Appendix A, based on Equation (16). Next, runs were filtered to only include the range of m values centered on $m = 0$ for which S_m remained sufficiently large. Using this *reduced* range, only clusters with ECI (entry in \overline{v}_m) that were nonzero more than 50% of the time were kept, forming the “selected” set. Finally, ECIs were calculated for the selected set of clusters at each m in the reduced range, with each m being treated independently. By averaging \overline{v}_m of the selected set of clusters over the reduced range of m , a final sparse set of ECIs was determined.

TABLE I. Simulation conditions for semi-grand canonical Monte Carlo simulations.

| Simulation | N_{pass} | τ_0 | $\Delta\tau$ | Δm |
|------------|-------------------|----------|------------------------|------------|
| I (2D) | 10,000 | 1.29 | -2.59×10^{-3} | 0.15 |
| II (2D) | 5,000 | 6.46 | -12.9×10^{-3} | 0.3 |
| III (3D) | 5,000 | 3.18 | -3.18×10^{-3} | 0.369 |

A. System I: NN and NNN 2D Triangular Lattice

The original and recovered clusters and ECIs for the 2D triangular lattice are given in Table II, with diagrams of the clusters shown in Figure 1a. The zero Kelvin formation energies of several structures, including the five ground states for this cluster expansion, are shown in Figure 1b, with the ordering of each ground state illustrated in Figure 1c. This set of clusters and ECIs produces a symmetric phase diagram with both first-order and continuous phase transitions as is evident in Figure 2a. Data for use in our algorithm was sampled over a wide range of chemical potentials and at temperatures from the region in Figure 2a bounded by blue, dashed lines.

TABLE II. Cluster Characteristics for the 2-Cluster 2D Triangular Lattice.

| Cluster i | Original v_i | Recovered v_i |
|-------------|----------------|-----------------------|
| 2 | 1 | 0.969 |
| 3 | 0.1 | 0.0966 |
| 12 | 0 | 1.26×10^{-3} |

The algorithm of Appendix A was applied to data generated over a range of chemical potentials, m , yielding

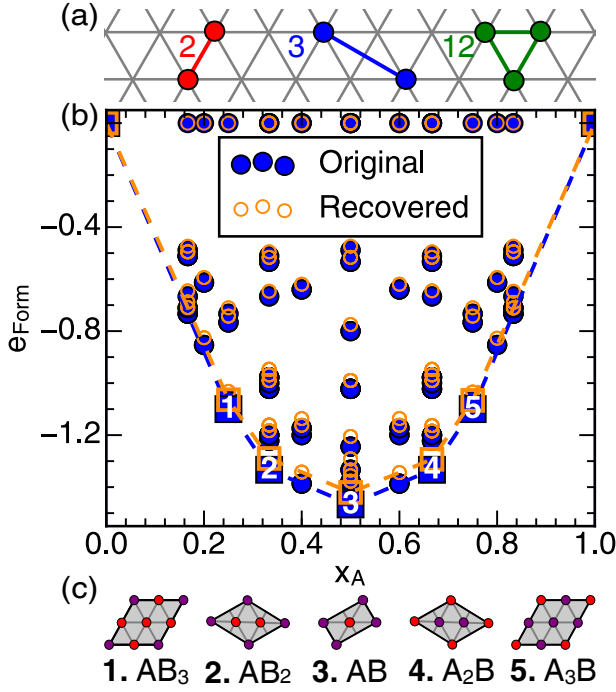


FIG. 1. (a) shows the two initial cluster prototypes used in our 2D triangular lattice (2 and 3), in addition to a third recovered cluster prototype (12). (b) shows the composition vs formation energy of a selection of configurations, in supercells containing up to 6 sites. Squares indicate ground states and are numbered to match (c). (c) shows schematic cells of the five ordered ground states. Red circles represent particle A, $\sigma_i = +1$, and purple circles represent particle B, $\sigma_i = -1$.

a sparse set of ECIs at each m . The region where the algorithm performs consistently was determined by the location of the first significant increase in the consistency score, S_m , (Equation (17)) surrounding $m = 0$, as can be seen in Figure 3a. The region of data then used to determine the final clusters and their ECIs is indicated by the orange (dashed) lines in Figures 3a and 2a, referred to as the “reliable zone”. The final ECIs were determined following two steps: first, the percentage of m values in the reliable zone in which each cluster was included in the cluster expansion was tallied. This percentage is presented in Figure 3b. Any clusters which appeared in half or more of the runs in the reliable zone were included in the final set of clusters. This final set of clusters was then used in a global regression over data collected at all chemical potential values m in the reliable zone. The final set of ECI are listed in Table II.

In addition to the nearest and next-nearest neighbor clusters (2 and 3, respectively), the algorithm also picked up the nearest-neighbor triplet (cluster 12). The recovered ECIs of clusters 2 and 3 are both within 5% of their original values, in addition to maintaining the 10:1 ratio present in the original cluster expansion. The nearest-neighbor triplet has a value nearly two orders-of-magnitude smaller than that of the next-nearest neighbor

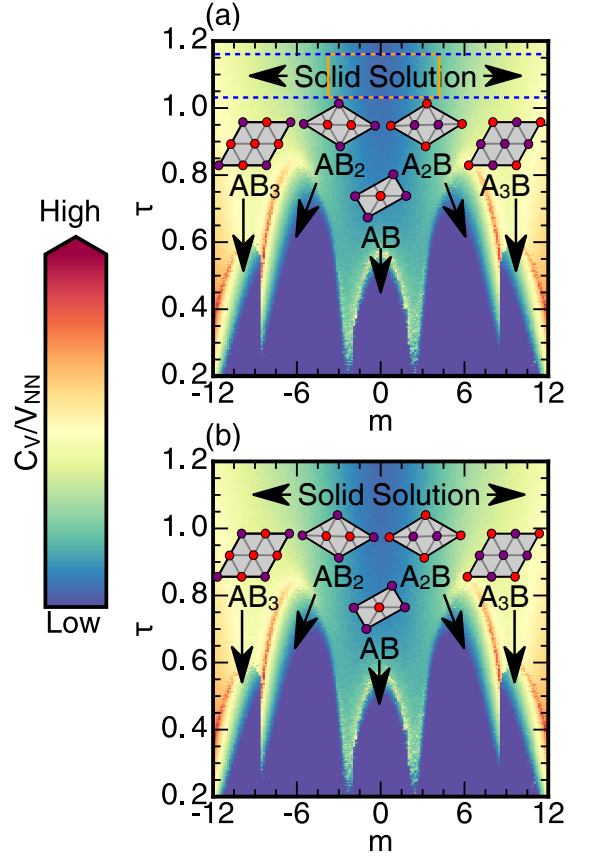


FIG. 2. Plots (a) and (b) show logarithmic heatmaps of the heat capacity C_V (scaled by V_{NN}), using the original and recovered ECIs, respectively. The approximate phase boundaries visible as sharp shifts in color, and appear at nearly identical locations in both phase maps. The blue dashed lines indicate the range of temperatures across which observations were taken, while the orange lines match those in Figure 3a.

ECI; its impact on any calculated energies is therefore negligible. This assertion is proved by both the zero Kelvin formation energies reproduced using the recovered ECIs in Figure 1b, and the shape and features of the phase diagram in Figure 2b. For this simple model cluster expansion, the algorithm of Appendix A has successfully recovered not only the correct ground states, but the correct phase behavior throughout all of phase space, while utilizing only a tiny fraction of the data available.

B. System II: 6-Cluster 2D Triangular Lattice

To examine a more complex cluster expansion for the triangular lattice, six clusters were chosen to represent a spread of cluster lengths and cluster sizes. The values of the ECIs were chosen randomly and are listed in Table III. Their corresponding clusters are shown in Figure 4a. Zero Kelvin formation energies for a selection of orderings on the triangular lattice, including the five

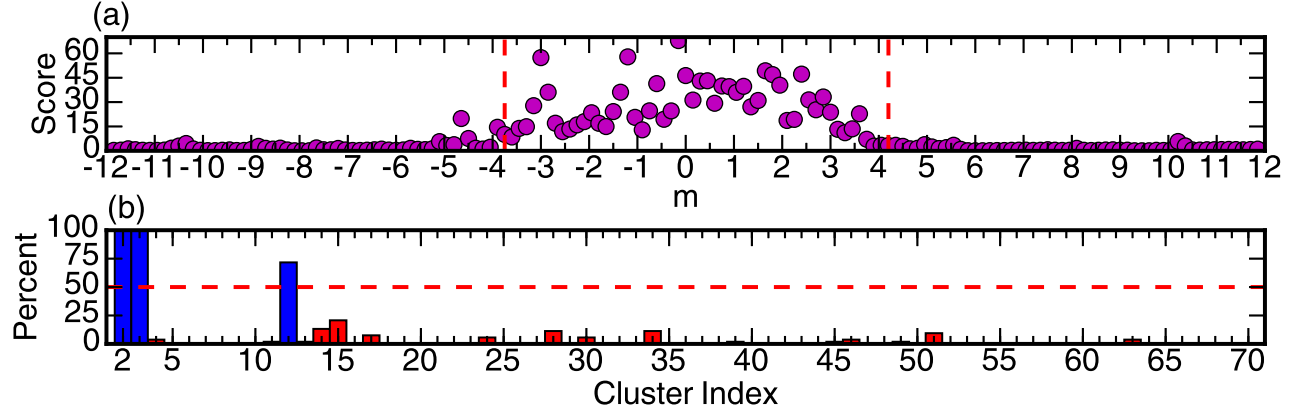


FIG. 3. (a) shows the consistency score (Equation 17) calculated at each chemical potential (purple dots) using the recovered clusters and ECIs found via our algorithm. Only data between the dashed orange lines was used for subsequent analysis. (b) shows the fraction of runs each cluster appeared in; only clusters above the cutoff ($\geq 50\%$, dashed red line) were utilized in the final regression step to determine ECIs.

ground states, are shown in Figure 4b, with orderings for each ground state illustrated in Figure 4c. The phase diagram for this cluster expansion is shown in Figure 5a and is asymmetric, exhibiting both first-order and continuous phase transitions. As before, only data at temperatures bounded by the two blue, dashed lines in Figure 5a was used in the algorithm of Appendix A to recover the ECI.

TABLE III. Cluster Characteristics for the 6-Cluster 2D Triangular Lattice.

| Cluster i | Original v_i | Recovered v_i |
|-------------|----------------|-----------------|
| 2 | 1 | 0.946 |
| 3 | 0.3 | 0.266 |
| 6 | 0.5 | 0.433 |
| 12 | 0.3 | 0.286 |
| 14 | 0 | -0.0109 |
| 15 | 0.5 | 0.450 |
| 47 | 0.3 | 0.287 |

Similar to system I described in Section IIIA, we calculated a consistency score for each chemical potential, and used an increase in the consistency score to bound the “reliable zone”. The scores and resulting boundaries are shown in Figure 6a. The cluster frequencies in this region are plotted in Figure 6b, with clusters selected more than 50% of the time utilized in the final series of regressions. In addition to the original clusters, the next-nearest-neighbor triplet (cluster 14) was picked up, with an ECI one order-of-magnitude smaller than the next-smallest ECI. All of the remaining ECIs recovered were within 15% of their original values, and 10% of their relative relationships to the nearest-neighbor ECI.

The recovered cluster expansion correctly reproduces the same ground states and formation energies (with a vertical offset) of the original cluster expansion, as shown

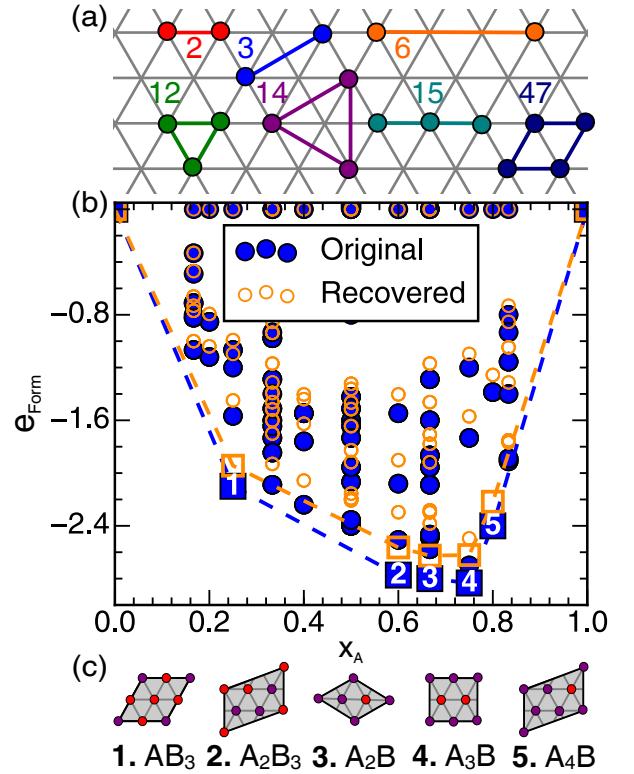


FIG. 4. (a) shows the six initial cluster prototypes (2, 3, 6, 12, 15, 47) used to generate data, as well as a spuriously-recovered cluster prototype (14). (b) shows the composition versus formation energy for all configurations in supercells containing up to 6 sites. Squares indicate ground states and are numbered to match (c). (c) shows schematic cells of the five ordered ground states. Red circles represent particle A , $\sigma_i = +1$, and purple circles represent particle B , $\sigma_i = -1$.

in Figure 4b. The calculated phase diagram of Figure 5b shows that the transition temperatures and the nature of

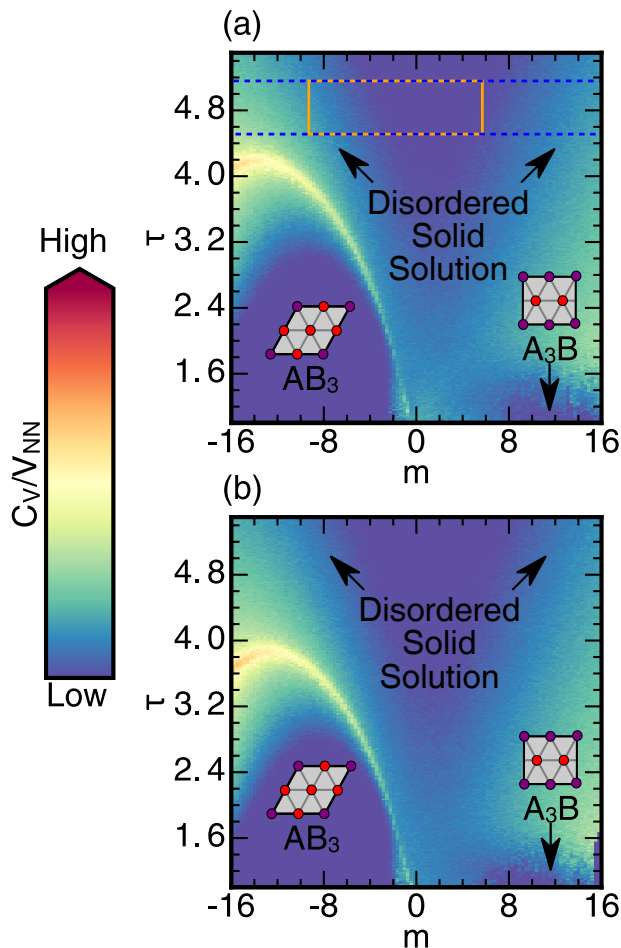


FIG. 5. Plots (a) and (b) show logarithmic heatmaps of the heat capacity C_V (scaled by V_{NN}), using the original and recovered ECIs, respectively. The approximate phase boundaries are visible as sharp shifts in color, and appear at nearly identical locations in both phase maps, save for a slight amount of scaling. The blue dashed lines indicate the range of temperatures across which observations were taken, while the orange lines match those in Figure 6a.

the transition (i.e., first-order versus continuous) are also faithfully reproduced across all of phase space. These results demonstrate the ability of the algorithm to recover a cluster expansion from high temperature data that correctly predicts phase stability over all of phase space.

C. System III: 3D FCC Lattice

The algorithm of Appendix A was also tested on a cluster expansion constructed by Z. Lu, *et al.*⁶³ to describe the Au-Cu binary alloy. The ECIs are given in Table IV and the clusters are illustrated in Figure 7a. The zero Kelvin formation energies of the L1₀ and L1₂ ground states, as well as of a number of other configurations, are shown in Figure 7b with orderings of the ground states illustrated in Figure 7c. Figure 8a shows

the phase diagram, exhibiting expected behavior akin to that experimentally observed for Au-Cu. As for model systems I and II, only data from temperatures between the two blue, dashed lines in Figure 8a was used to recover a cluster expansion.

TABLE IV. Cluster Characteristics for the Au-Cu FCC Lattice.

| Cluster i | Original v_i | Recovered v_i |
|-------------|----------------|-----------------|
| 2 | 1 | 0.947 |
| 10 | 0 | 0.0170 |
| 3 | 0.0224 | 0 |
| 12 | 0.0622 | 0.0764 |
| 4 | 0.0576 | 0 |
| 32 | 0.0129 | 0 |
| 5 | 0.0157 | 0 |

The “reliable zone” of chemical potentials was again determined using the consistency score (Figure 9a) and the final choice of clusters was determined by the frequency with which they were picked up by the algorithm for each chemical potential within this zone (Figure 9b). In this case, a *reduced* set of clusters was recovered, excluding the 4-body cluster and many of the 2-body clusters, but including a new longer-range pair interaction, shown as cluster 10 in Figure 7a. This new set of clusters correctly reproduces the ground states as well as the on-the-hull degenerate configurations identified using the original cluster expansion as can be seen in Figure 7b. Monte Carlo simulations applied to the recovered cluster expansion also faithfully reproduces the phase behavior of the system to within a scaling factor, visible in Figure 8b.

While the recovered cluster expansion in this example differs qualitatively from the original one in terms of the number and types of clusters, it nevertheless correctly reproduces the ground states and the finite temperature phase diagram. Equation (16) guarantees a deterministic set of clusters and ECIs, but it does not necessarily guarantee the *same* set of clusters will be picked up as those used to generate the high temperature data. By using the MAXENT method, we bias our recovery towards specific sets of solutions. This example illustrates that multiple sets of clusters and ECIs can generate the same phase behavior. Therefore, while the original set of clusters and ECIs can produce the phase diagram in Figure 8a, we have recovered another solution with qualitatively the same phase behavior.

IV. DISCUSSION

We have introduced a method to parameterize an atomistic Hamiltonian that is capable of accurately predicting both the thermodynamic ground states as well as the full phase diagram at finite temperature using

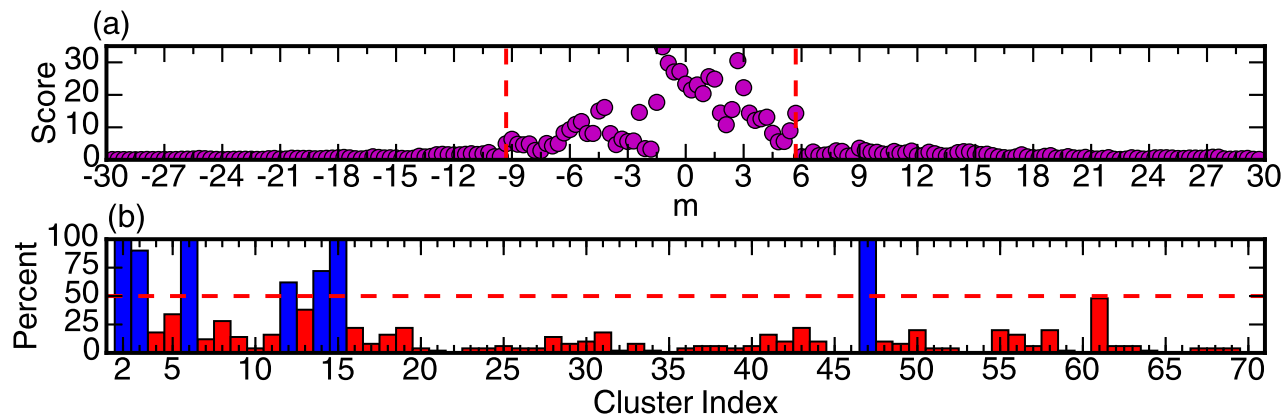


FIG. 6. (a) shows the score (Equation (17)) calculated at each chemical potential (purple dots) using the recovered clusters and ECIs found via our algorithm. Only data from between the dashed orange lines was used for subsequent analysis. (b) shows the fraction of runs each cluster appeared in; only clusters above the cutoff ($\geq 50\%$, red dashed line) were used in the final regression to determine the ECIs.

only information about the disordered state. We demonstrated the approach for binary alloys modelled with cluster expansion Hamiltonians, which express the dependence of the energy of a multi-component crystal as a linear expansion of cluster basis functions. The foundation of the approach rests on a thermodynamic interpretation of the cluster expansion formalism: extensive cluster basis functions, Φ_α , and their corresponding effective cluster interaction (ECI) coefficients, V_α , form conjugate pairs like any other set of thermodynamic variables. Such an interpretation reveals the existence of Maxwell relations that can be converted to a set of equations (Equation (11)) relating two sets of measurable quantities (i.e., the covariance between pairs of extensive cluster functions and the temperature derivative of the ensemble averages of extensive cluster functions) to the unknown ECI of a cluster expansion.

Cluster expansions parameterized from first principles tend to be sparse, and not require the complete set of basis functions. We have shown that a thermodynamic interpretation of the ECIs also implies a free energy-like function, Equation (13). This free energy-like function has a maximum which corresponds to a specified set of basis functions to be retained in a truncated expansion consistent with the *observed* averages of the cluster basis functions. This follows from Jaynes’ maximum (information) entropy or MAXENT approach.

The two properties described in Equations (11) and (16) emerge from a thermodynamic interpretation of the cluster expansion formalism. These thermodynamic features motivate and support an iterative algorithm for the parameterization of an effective Hamiltonian to high temperature observations. The final step relies on a regression model to invert Equation (11). However, since the measured system is exactly determined (one linear relation and one unknown for each cluster basis function of a cluster expansion), direct inversion of Equation (11)

becomes both numerically unstable and computationally intractable as the number of cluster basis functions becomes exceedingly large in the thermodynamic limit. Furthermore, most multi-component solids can be accurately described with a sparse cluster expansion where only a small subset of the ECI are non-zero. Hence, ordinary least squares is not a suitable method for regression, even if only considering the first n rows and first $m < n$ columns of Equation (11). Furthermore, we are also prohibited from sparsity-preserving techniques such as LASSO⁵⁹ due to the nature of both the regressors (the covariances) and the observed variable (the change in extensive cluster functions with temperature). It is in this context that an initial step involving a maximization of the free energy, Equation (13), using an approximation for the disordered state, Equation (16), can guide the selection of a sparse truncated cluster expansion (i.e., a sparse set of non zero ECI). Iteration between inverting a sparse form of Equation (11) and maximizing Equation (13) then leads to a Hamiltonian that is consistent with high temperature measurements of the disordered state. As our three examples illustrated, the Hamiltonians parameterized this way are capable of reproducing the ground state orderings as well as the topology of finite temperature phase diagrams with remarkable accuracy.

The approach introduced here differs from conventional inverse Monte Carlo schemes^{58,66–69}, which seek to recover interaction parameters of a Hamiltonian from measures of average cluster functions. While inverse Monte Carlo methods can generate similar Hamiltonians as the approach introduced here, they require a new round of Monte Carlo simulations for each step in the gradient descent towards the “correct” ECIs. Furthermore, inverse Monte Carlo methods provide no proscription as to *which* cluster basis functions to query, leading to instability of the solution when numerous spurious cluster functions are considered simultaneously for the case of a

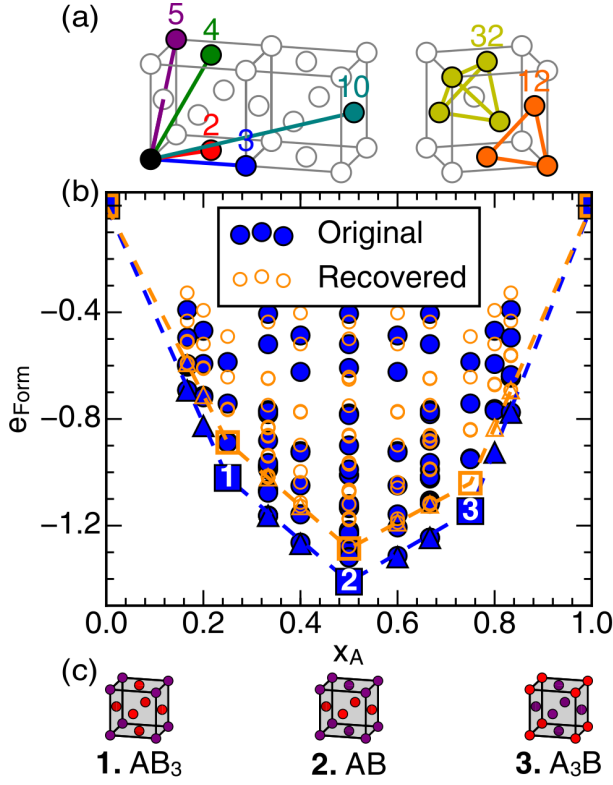


FIG. 7. (a) shows the six initial cluster prototypes (2, 3, 4, 5, 12, 32) used to generate data, as well as the newly recovered cluster prototype (10). (b) shows the composition versus formation energy for all configurations in supercells containing up to 6 sites, and selected supercells containing up to 8 sites. Squares indicate ground states and are numbered to match (c), triangles indicate degenerate configurations that lie along, but do not deform, the common tangent between adjacent ground states. (c) shows schematic cells of the three ordered ground states. Red circles represent particle A , $\sigma_i = +1$, and purple circles represent particle B , $\sigma_i = -1$.

sparse ground-truth. The approach of this work, in contrast, does not require iteration with Monte Carlo and relies on an agnostic approach in the selection of relevant cluster basis functions. In fact, the step relying on Equation (13) can also be incorporated in conventional inverse Monte Carlo schemes as a way of cluster function selection.

We have said much about “experimental observables” without yet discussing how the $\langle \Phi \rangle$ and $\text{cov}[\Phi, \Phi]$ may actually be obtained. Ultimately, the extensive cluster functions are merely the products of site occupation variables, and so if provided with exact atomic data, one would map each site onto a lattice, assign a spin variable, and be able to directly calculate Φ . The averages and covariances of the clusters can then be calculated by sub-dividing a sufficiently large observation into N smaller observations and taking averages and covariances across this collection of observations. This sort of exact atomic information is available via atom probe tomography, which can yield observations with volumes on the

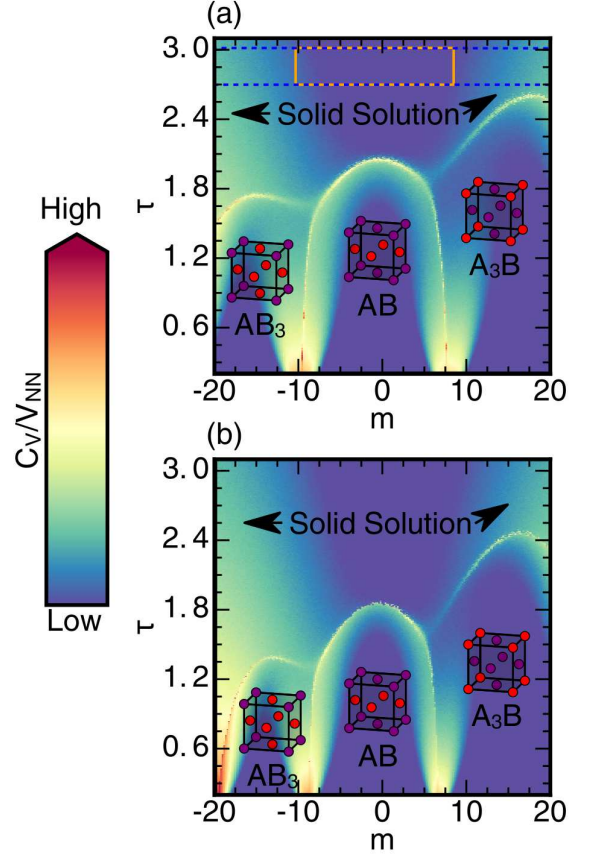


FIG. 8. Plots (a) and (b) show logarithmic heatmaps of the heat capacity C_V (scaled by V_{NN}), using the original and recovered ECIs, respectively. The approximate phase boundaries are visible as sharp shifts in color, and appear at nearly identical locations in both phase maps, save for a slight amount of scaling. The blue dashed lines indicate the range of temperatures across which observations were taken, while the orange lines match those in Figure 9a.

order of 10^6 nm^{370} or 10^8 unit cells. High-angle annular dark-field imaging (HAADF-STEM) can also provide atomic-scale resolution of a sample as well, and by varying the depth of focus, a 3D image can be obtained over a comparable volume⁷¹. Atom probe and HAADF-STEM do not provide 100% coverage of the volumes they query, but each provides a sufficient overabundance of information as to allow for some guesses at the unknown zones. Less directly, information on pair correlations can be obtained via techniques such as x-ray and electron diffraction, the covariances of the pair cluster functions using fluctuation microscopy⁷², and short-range pair and multi-body terms using nuclear magnetic resonance (NMR) multiple-quantum experiments⁷³. However, rather than use diffraction or NMR techniques directly, it is likely these could be used to supplement any interpretation of atom probe or HAADF-STEM analysis, providing better guesses at information that may be missing.

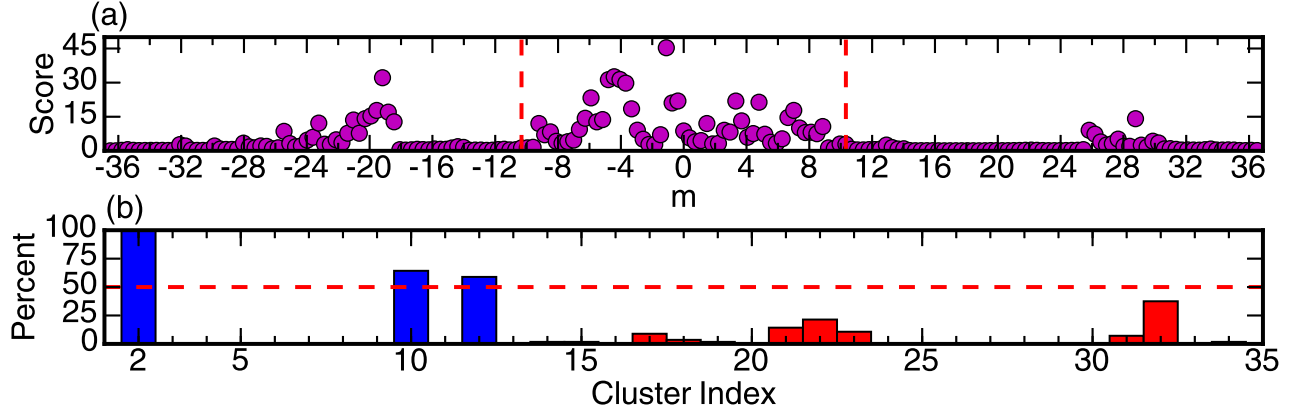


FIG. 9. (a) shows the score (Equation (17)) calculated at each chemical potential (purple dots) using the recovered clusters and ECIs found via our algorithm. Only data from between the dashed, orange lines was used for subsequent analysis. (b) shows the fraction of runs each cluster appeared in; only clusters above the cutoff ($\geq 50\%$, red dashed line) were used in the final regression to determine the ECIs.

While the approach introduced here has been developed in the context of a binary alloy Hamiltonian, it can be used to invert any effective Hamiltonian that is expressed as a linear expansion of basis functions of relevant degrees of freedom. The linear expansion coefficients that measure the weight of a particular basis function in the effective Hamiltonian can again be interpreted as a thermodynamic variable. Equations similar to (11) and (16) can then be derived that, through an iterative procedure, enable the parameterization of interaction coefficients using measurements in a high temperature phase. The types of Hamiltonians that can be analyzed in this manner include multi-component (i.e., ternary, quaternary, etc.) cluster expansions, spin-cluster expansions describing non-collinear magnetic solids¹⁶ and lattice dynamical Hamiltonians in the harmonic approximation and beyond.^{12,19,29,46,74}

V. CONCLUSION

We have developed a new method to recover relevant interaction coefficients of effective Hamiltonians from experimentally measurable quantities. By careful examination and manipulation of the free energy, a simple mathematical relationship between fluctuations of extensive

cluster functions and their related interaction coefficients emerges. The numerical instability of this equation is solved by the development of a secondary criterion, based on the principle of maximum entropy as put forth by Jaynes. Using a single pass through the space of basis functions of the Hamiltonian, we recover a unique solution in polynomial time. The method has been tested in multiple *in-silico* experiments, and faithfully reproduced both the original thermodynamic ground states and the full phase diagrams of each of our simulated systems.

ACKNOWLEDGMENTS

Anirudh Natarajan and John Goiri for many helpful discussions and assistance with performing calculations and Dr. Paul Weakliem for help with computational facilities. Elizabeth Decolvenaere is supported by the MRSEC Program of the Natural Science Foundation under Award No. DMR-1121053. Professor Van der Ven is grateful for funding provided by the Office of Naval Research under Grant No. N00014-12-1-0013 (project manager William Mullins). Simulations were performed using resources from the Center for Scientific Computing in the CNSI and MRL, funded by NSF MRSEC (DMR-1121053), NSF CNS-0960316, and Hewlett Packard.

* avdv@engineering.ucsb.edu

¹ A. Pasquarello, M. S. Hybertsen, and R. Car, Lett. to Nat. **396**, 58 (1998).

² H. R. Rüter and R. Redmer, Phys. Rev. Lett. **112**, 145007 (2014).

³ Y. Wang, W. D. Richards, S. P. Ong, L. J. Miara, J. C. Kim, Y. Mo, and G. Ceder, Nat Mater **14**, 1 (2015).

⁴ Y. Mishin, M. Asta, and J. Li, Acta Mater. **58**, 1117 (2010).

⁵ A. P. Thompson, L. P. Swiler, C. R. Trott, S. M. Foiles, and G. J. Tucker, J. Comput. Phys. **285**, 316 (2015).

⁶ M. S. Daw and M. I. Baskes, Phys. Rev. B **29**, 6443 (1984).

⁷ J. M. Sanchez, F. Ducastelle, and D. Gratias, Phys. A Stat. Mech. its Appl. **128**, 334 (1984).

- ⁸ J. Tersoff, Phys. Rev. B **37**, 6991 (1988).
- ⁹ J. M. Sanchez, Phys. Rev. B **48**, 14013 (1993).
- ¹⁰ W. Zhong, D. Vanderbilt, and K. M. Rabe, Phys. Rev. Lett. **73**, 1861 (1994), arXiv:9406049 [cond-mat].
- ¹¹ K. M. Rabe and U. V. Waghmare, Phys. Rev. B **52**, 13236 (1995), arXiv:9411006 [mtrl-th].
- ¹² W. Zhong, D. Vanderbilt, and K. M. Rabe, Phys. Rev. B **52**, 6301 (1995), arXiv:9502004 [mtrl-th].
- ¹³ T. S. Bush, J. D. Gale, C. R. A. Catlow, and P. D. Battle, J. Mater. Chem. **4**, 831 (1994).
- ¹⁴ D. de Fontaine, Solid State Phys. **47**, 33 (1994).
- ¹⁵ A. C. T. van Duin, S. Dasgupta, F. Lorant, and G. W. A., J. Phys. Chem. A **105**, 9396 (2001).
- ¹⁶ R. Drautz and M. Fähnle, Phys. Rev. B **69**, 1 (2004).
- ¹⁷ R. Drautz, X. W. Zhou, D. A. Murdick, B. Gillespie, H. N. G. Wadley, and D. G. Pettifor, "Analytic bond-order potentials for modelling the growth of semiconductor thin films," Walle and G. Ceder, J. Phase Equilibria **23**, 348 (2002), arXiv:0201511 [cond-mat].
- ¹⁸ J. Bhattacharya, A. der Ven, and A. Van der Ven, Acta Mater. **56**, 4226 (2008).
- ¹⁹ J. C. Thomas and A. Van der Ven, Phys. Rev. B **88**, 214111 (2013).
- ²⁰ J. C. Wojdel, P. Hermet, M. P. Ljungberg, P. Ghosez, and J. Ñíguez, J. Phys. Condens. Matter **25**, 305401 (2013), arXiv:arXiv:1301.5731v1.
- ²¹ T. P. Senftle, S. Hong, M. M. Islam, S. B. Kylasa, Y. Zheng, Y. K. Shin, C. Junkermeier, R. Engel-Herbert, M. J. Janik, H. M. Aktulga, T. Verstraelen, A. Grama, and A. C. T. van Duin, npj Comput. Mater. **2**, 15011 (2016).
- ²² A. Zunger, L. G. Wang, G. L. W. Hart, and M. Sanati, Model. Simul. Mater. Sci. Eng. **10**, 685 (2002).
- ²³ B. P. Burton and A. van de Walle, Calphad **39**, 97 (2012).
- ²⁴ B. Puchala and A. Van der Ven, Phys. Rev. B **88**, 1 (2013).
- ²⁵ D. Chang, M. H. Chen, and A. Van Der Ven, Chem. Mater. **27**, 7593 (2015).
- ²⁶ A. R. Natarajan, E. L. Solomon, B. Puchala, E. A. Marquis, and A. Van der Ven, Acta Mater. **108**, 367 (2016).
- ²⁷ B. Sadigh and P. Erhart, Phys. Rev. B **86**, 134204 (2012).
- ²⁸ A. Van der Ven, J. C. Thomas, Q. Xu, B. Swoboda, and D. Morgan, Phys. Rev. B **78**, 104306 (2008).
- ²⁹ A. Van der Ven, J. C. Thomas, Q. Xu, and J. Bhattacharya, Math. Comput. Simul. **80**, 1393 (2010).
- ³⁰ J. Bhattacharya and A. Van Der Ven, Phys. Rev. B **83**, 144302 (2011).
- ³¹ V. Vaithyanathan, C. Wolverton, and L. Chen, Acta Mater. **52**, 2973 (2004).
- ³² D. B. Laks, L. G. Ferreira, S. Froyen, and A. Zunger, Phys. Rev. B **46**, 12587 (1992), arXiv:arXiv:1011.1669v3.
- ³³ A. Zunger, in *Statics Dyn. Alloy Phase Trans- Form.* (Springer US, 1994) p. 361.
- ³⁴ P. D. Tepesch, G. D. Garbulsky, and G. Ceder, Phys. Rev. Lett. **74**, 2272 (1995).
- ³⁵ V. Blum and A. Zunger, Phys. Rev. B **70**, 155108 (2004).
- ³⁶ A. Van der Ven and G. Ceder, Phys. Rev. B **71**, 054102 (2005).
- ³⁷ F. Zhou, T. Maxisch, and G. Ceder, Phys. Rev. Lett. **97**, 155704 (2006), arXiv:0612163 [cond-mat].
- ³⁸ A. Seko, K. Yuge, F. Oba, A. Kuwabara, and I. Tanaka, Phys. Rev. B **73**, 184117 (2006).
- ³⁹ B. P. Burton, A. Van De Walle, and U. Kattner, J. Appl. Phys. **100**, 113528 (2006).
- ⁴⁰ O. Levy, G. L. W. Hart, and S. Curtarolo, Phys. Rev. B **81**, 174106 (2010), arXiv:1002.2822.
- ⁴¹ K. Yuge and R. Okawa, Intermetallics **44**, 60 (2014).
- ⁴² A. Belak and A. Van der Ven, Phys. Rev. B **224109**, 1 (2015).
- ⁴³ M.-H. Chen, B. Puchala, and A. Van der Ven, Calphad **51**, 292 (2015).
- ⁴⁴ E. Decolvenaere, M. J. Gordon, and A. Van der Ven, Phys. Rev. B **92**, 085119 (2015), arXiv:1508.02737.
- ⁴⁵ J. Goiri and A. Van der Ven, 10.1103/PhysRevB.00.004100.
- ⁴⁶ J. C. Thomas and A. Van Der Ven, Phys. Rev. B **90**, 224105 (2014).
- ⁴⁷ J. W. D. Connolly and A. R. Williams, Phys. Rev. B **27**, 5169 (1983).
- ⁴⁸ G. Ceder, G. D. Garbulsky, and P. D. Tepesch, Phys. Rev. B **51**, 11257 (1995).
- ⁴⁹ "Conductivity in thin films," Walle and G. Ceder, J. Phase Equilibria **23**, 348 (2002), arXiv:0201511 [cond-mat].
- ⁵⁰ G. L. W. Hart, V. Blum, M. J. Walorski, and A. Zunger, Nat. Mater. **4**, 391 (2005).
- ⁵¹ T. Mueller and G. Ceder, Phys. Rev. B **80**, 024103 (2009).
- ⁵² T. Mueller and G. Ceder, Phys. Rev. B **82**, 184107 (2010).
- ⁵³ E. Cockayne and A. Van De Walle, Phys. Rev. B **81**, 012104 (2010), arXiv:0908.0659.
- ⁵⁴ L. J. Nelson, G. L. W. Hart, F. Zhou, and V. Ozoliš, Phys. Rev. B **87**, 035125 (2013), arXiv:1208.0030v2.
- ⁵⁵ L. J. Nelson, V. Ozoliš, C. S. Reese, F. Zhou, and G. L. W. Hart, Phys. Rev. B **88**, 155105 (2013), arXiv:1307.2938.
- ⁵⁶ J. Kristensen and N. J. Zabaras, Comput. Phys. Commun. **185**, 2885 (2014).
- ⁵⁷ A. van de Walle and M. Asta, Model. Simul. Mater. Sc. **10**, 521 (2002).
- ⁵⁸ F. Livet, Acta Met. **35**, 2915 (1987).
- ⁵⁹ R. Tibshirani, J. R. Stat. Soc. Ser. B (Statistical Methodol. **73**, 273 (2011), arXiv:11/73273 [13697412].
- ⁶⁰ B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, H. Ishwaran, K. Knight, J. M. Loubes, P. Massart, D. Madi-gan, G. Ridgeway, S. Rosset, J. I. Zhu, R. A. Stine, B. A. Turlach, S. Weisberg, T. Hastie, I. Johnstone, and R. Tibshirani, Ann. Stat. **32**, 407 (2004), arXiv:0406456 [math].
- ⁶¹ E. T. Jaynes, Phys. Rev. **106**, 620 (1957).
- ⁶² James Glosli; Michael Plischke, Can. J. Phys. **61**, 1515 (1983).
- ⁶³ Z. W. Lu, S.-H. Wei, A. Zunger, S. Frota-Pessoa, and L. G. Ferreira, Phys. Rev. B **44**, 512 (1991).
- ⁶⁴ B. Puchala, M. Radin, N. S. H. Gunda, J. Goiri, L. Decolvenaere, A. R. Natarajan, and J. C. Thomas, (2016), 10.5281/ZENODO.60142.
- ⁶⁵ Specifically, the number of configurations C is now $C = 2^N$, and the point-term ECI changes from V_1 to $V'_1 = V_1 + \frac{\mu}{2}$. This transformation is achieved by examining the argument to the exponential in the semigrand canonical ensemble: $\frac{-V \cdot \Phi(\bar{\sigma}) + \mu N_A}{k_B T}$, and noting that, for our choice of basis set, $N_A = \frac{\Phi_1(\bar{\sigma}) + N}{2}$. We can then collect $\Phi_1(\bar{\sigma})(V_1 + \frac{\mu}{2}) = \Phi_1(\bar{\sigma})V'$ and bring the remaining $\frac{\mu N}{2}$ out of the exponential, and indeed out of any outer sum over microstates, entirely as the number of sites N is unchanging.
- ⁶⁶ V. Gerold and J. Kern, Acta Metall. **35**, 393 (1987).
- ⁶⁷ K. Mosegaard and A. Tarantola, J. Geophys. Res. **100**, 12431 (1995).

- ⁶⁸ K. Mosegaard and M. Sambridge, Inverse Probl. **18**, R29 (2002).
- ⁶⁹ J. Albert and R. H. Swendsen, Phys. Procedia **57**, 99 (2014).
- ⁷⁰ A. Cerezo, P. H. Clifton, M. J. Galtrey, C. J. Humphreys, T. F. Kelly, D. J. Larson, S. Lozano-Perez, E. A. Marquis, R. A. Oliver, G. Sha, K. Thompson, M. Zandbergen, and R. L. Alvis, Mater. Today **10**, 36 (2007).
- ⁷¹ K. Sohlberg, T. J. Pennycook, W. Zhou, and S. J. Pennycook, Phys. Chem. Chem. Phys. **17**, 3982 (2015).
- ⁷² M. M. J. Treacy, J. M. Gibson, L. Fan, D. J. Paterson, and I. McNulty, Reports Prog. Phys. **68**, 2899 (2005).
- ⁷³ S. Cadars, A. Lesage, N. Hedin, B. F. Chmelka, and L. Emsley, J. Phys. Chem. B **110**, 16982 (2006).
- ⁷⁴ G. D. Garbulsky and G. Ceder, Phys. Rev. B **49**, 6327 (1994).
- ⁷⁵ A *geometric cluster* refers to the set of sites on a lattice included in the definition of a cluster function, but *not* which basis function of the site is being used.
- ⁷⁶ Ridge regression is only used in the *selection* step, not in the for the final ECIs when A_s has been fully determined; the motivation for this choice is described in Appendix C.
- ⁷⁷ F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011), arXiv:1201.0490.

Appendix A: Selection Algorithm

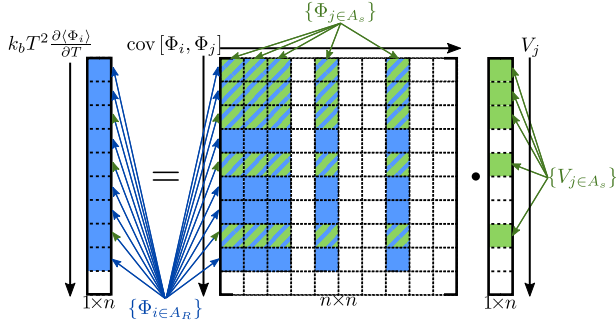


FIG. 10. Schematic version of the regression scheme, Equation (11), indicating which columns are selected by Equation (16) (the set A_s , in green) and which rows are included by the cluster radius cutoff (the set A_R , in blue). Selecting only rows that represent extensive cluster functions in A_s results in large quantities of data going unused. By using A_R , i.e., the set of geometric clusters at or below the cluster radius of the largest geometric cluster in A_s , a significantly larger portion of the available observations can be used.

The approach of our algorithm is to invert Equation (11) via regression, using Equation (16) to select which ECIs will be allowed to be nonzero. Additionally, we wish to restrict the range of measurements, i.e., $k_b T \frac{\partial \langle \Phi_i \rangle}{\partial T}$, utilized in our regression. Figure 10 schematically illustrated the set-up of the regression, indicating which ECIs have been selected (in green) and which

measurements are being utilized (in blue). Let some selected set of ECIs be indexed by $A_s = \{\alpha_0, \alpha_1, \dots\}$, such that our model coefficients are $\bar{V}_{A_s} = \{V_{i \in A_s}\}$. Then, let some utilized set of measurements be indexed by $A_R = \{\beta_0, \beta_1, \dots\}$, such that our predicted outputs are $\bar{y}_{A_R} = \{k_b \frac{\partial \langle \Phi_{j \in A_R} \rangle}{\partial T}\}$. Then, the columns utilized in our design matrix, \mathbf{X} , must be A_s and the rows utilized must be A_R , such that $\mathbf{X} = \text{cov}[\{\Phi_{j \in A_R}\}, \{\Phi_{i \in A_s}\}]$. These definitions will be utilized extensively in the description of our algorithm, below. Our selection-and-regression algorithm runs in polynomial time, and requires no additional data or information beyond the same information needed for Equation (11). The algorithm as implemented in section III is outlined below:

1. Form the set of extensive cluster function indices A for which there exists data, up to a cut-off radius r , sorted by increasing geometric cluster⁷⁵ size (first by cluster radius, then by the number of sites in the cluster):

$$A = \{\alpha_0, \alpha_1, \dots, \alpha_n\}.$$

2. Initialize the list of selected ECIs A_s by selecting the indices associated with the empty, point and pair clusters from A :

$$A_s = \{\alpha_0, \alpha_1, \alpha_2\}.$$

3. Let the cluster radius $r(\alpha_i)$ be the longest distance between any two sites in α_i , and define R_s be the largest cluster radius of geometric clusters represented in A_s :

$$R_s = \max[\{r(\alpha_i) : \forall \alpha_i \in A_s\}]$$

4. Form the set of indices affiliated with geometric clusters as small as, or smaller than, R_s . This set defines the extensive cluster functions with measurements that we presume to be dominated by signal, rather than noise:

$$A_R = \{\alpha_i : \forall \alpha_i \in A \text{ if } r(\alpha_i) \leq R_s\}$$

5. Using ridge regression⁷⁶ (with regularization parameter γ), calculate the ECIs for A_s :

$$\bar{V}_{A_s} = (\mathbf{X} \mathbf{X}^\top + \gamma \mathbf{I})^{-1} \mathbf{X} \cdot \bar{y}_{A_R}$$

6. From A , select the next index α_j which has not yet been examined, and form the set A_j :

$$A_j = A_s + \{\alpha_j\}.$$

7. Using ridge regression⁷⁵ (with regularization parameter γ), calculate the ECIs for A_j :

$$\bar{V}_{A_j} = k_b T^2 (\mathbf{X} \mathbf{X}^\top + \gamma \mathbf{I})^{-1} \mathbf{X} \cdot \bar{\mathbf{y}}_{A_R}$$

8. Use Equation (16) to calculate $\Delta \Upsilon$:

$$\Delta_j \Upsilon(T, M, N_A, \bar{\Phi}_{\text{obs}}) = \sum_{\alpha \in A_s} \left(\bar{\Phi}_0^\alpha - \bar{\Phi}_{\text{obs}}^\alpha \right) \left(V_{A_j}^\alpha - V_{A_s}^\alpha \right)$$

9. If $\Delta_j \Upsilon > 0$, $A_s = A_j$, otherwise A_s remains unchanged.
10. Return to step 3, until there exist no indices in A which have not been examined.
11. Using the final set of selected A_s determine the ECIs. Calculate R_s and construct A_R as in steps 3 and 4, respectively, and build $\mathbf{X} = \text{cov}[\{\Phi_{j \in A_R}\}, \{\Phi_{i \in A_s}\}]$. Solve for \bar{V}_{A_s} using ordinary least squares regression:

$$\bar{V}_{A_s} = k_b T^2 (\mathbf{X} \mathbf{X}^\top)^{-1} \mathbf{X} \cdot \bar{\mathbf{y}}_{A_R}$$

Appendix B: The Effects of Order and Too Much Data

Our selection algorithm provides a unique set of clusters that can describe the observed data; the members of the set do *not* rely on the order in which clusters are considered. As the nature of Equation (14) does not depend on which subspace of clusters is chosen (the covariance matrix is, regardless, semipositive definite), results for $\Delta \Upsilon_j$ will remain correct in sign even if some clusters have previously been included (or excluded) incorrectly. However, if the measurements (rows) utilized in our regression are dominated by noise, rather than signal, the results of the regression behave erratically. Specifically, we have found that if we include *all* measurements available to us (limited only by when we have chosen to cease

enumerating new extensive cluster functions to measure), the results of our algorithm, and in fact, of any regression (even when selecting the set of ECIs used in the underlying Hamiltonian we were trying to recover), were divergent. In this way, the values entering Equation 16 are then no longer representative of the data, making our scheme (and any scheme) meaningless. Therefore, addition to selecting which ECIs to include in the fit, we have also chosen to restrict which measurements we utilize in our regression; these choices are described in the previous section and illustrated in Figure 10.

Appendix C: Motivation for Ridge Regression in Selection

When performing extensive cluster function selection, we add a penalty term to A proportional to the l^2 -norm of the ECIs to guarantee numerical stability of both Equation (11) and (16) (i.e., ridge regression). The regularization parameter γ is chosen as part of the Bayesian ridge regression scheme as implemented in scikit-learn⁷⁷. By using ridge regression, we are implicitly assuming that the ECIs are Gaussian distributed with a mean of 0.

Our second derivatives from Equation (14) are now:

$$\frac{\partial^2 \Upsilon}{\partial \bar{V}^2} = -\frac{\text{cov}[\bar{\Phi}, \bar{\Phi}]}{k_b T} - \frac{2\gamma}{k_b T} < 0. \quad (\text{C1})$$

For $\gamma > 0$, this guarantees that A (and thus Υ) is negative definite (and has a unique maximum). In the regression portion of the selection step, the addition of a regularization term ensures that the modified covariance matrix has no (near-)zero eigenvalues, preserving the numerical stability of the solution. The trade-off, in the form of (uniform) shrinkage, is that the ECIs recovered are slightly smaller than their “true” values. Shrinkage of the ECIs provides no penalty in the selection stage as long as the sign of the ECIs is preserved. During the final regression after selection has been performed, only ordinary least squares regression is used to avoid solution bias and shrinkage of the ECIs.